AMIA Annu Symp Proc 2021:(submitted)
draft material -- please do not cite

Reproducibility in biomedical natural language processing: A FAIR approach to what we need to know

KB Cohen, PhD[1], A Ripple, MLS[2], A Ben Abacha, PhD[2], O Bodenreider, MD, PhD[2], O Hargraves, BA[3], K Verspoor, PhD[4], P Zweigenbaum, PhD[5], D Demner-Fushman, MD, PhD[2]
[1]U. Colorado School of Medicine, Denver, CO, USA; [2]U.S. National Library of Medicine, NIH, Bethesda, MD, USA; [3]U. Colorado, Boulder, CO, USA; [4]RMIT University, Australia; [5]U. Paris-Saclay, CNRS, LISN, France

Introduction

The philosopher of science Hans Radder put it well: in order to know whether or not an experiment has been reproduced, we first need to know what was actually done[1]. But, for natural language processing experiments, case studies have demonstrated that it is not always intuitively obvious what exactly needs to be reported[2]. Furthermore, work by Olorisade et al.[3] has demonstrated that this is not always clear even for a relatively constrained subfield, such as text mining research. When we begin to think about the fundamental issue of generalizability in reproducibility, the question becomes even more complicated, and the answers probably much more nuanced[4].

To address the basic question of *what was actually done in this natural language processing experiment?*, we take an approach based in the FAIR Principles for managing the products of scientific research[5]. Although there are alternatives, it is a good choice because of its wide acceptance in biomedical research. We propose here a two-part schema for representing a natural language processing research article as a collection of metadata. The first part models a research article itself. The approach to this is frame-based and is inspired by previous work in the domains of biomedical research publishing[6–8] and of standards for the reporting of experiments[9–11]. The second part is a list of values for that frame. Because community consensus is essential to the adoption of any such representation of scientific work[5], our overall approach includes a significant amount of solicitation of feedback from a diverse cross-section of the natural language processing community. Additionally, we tested the coverage of the ontology using frequency-based methods applied to the language processing and text mining literature from two relatively distinct communities—biomedical text mining, and the Association for Computational Linguistics family of conferences.

Materials and Methods

The representation of a research article consists of a frame with values for the following four items:

1. Topic: what is the paper primarily *about*?
2. Method: what was *done*?
3. Data: what kind of *material* was used?
4. Evaluation: *how* was the work evaluated, question answered, or hypothesis tested?

The entities needed to describe these four aspects can be organized into an ontology. The ontology is structured by the typical relations, i.e. is-a and has-part, and a few additional ones.

In order to minimize subjectivity, the first draft of the ontology was constructed based on the indexes and tables of contents of popular language processing textbooks. Definitions were taken from open-source materials, including the primary literature and Wikipedia, and reviewed by a lexicographer. The overall model of research articles, as well as the ontology for describing them, was evaluated in two ways: by quantitative comparison to frequency and terminological analyses of the literature, and by solicitation of feedback from researchers in the field.

Quantitative evaluation: We analyzed over 9,000 PubMed-indexed natural language processing and text mining research articles. The Sketch Engine terminology extraction tools[12] generated a silver standard for evaluation of coverage.

Expert feedback: We did initial annotations of the complete sets of PubMed-indexed publications of several authors. We then met with them individually, and they corrected the metadata that we assigned to their research articles.

Results

The ontology currently contains 390 unique concepts and several relation types. Table 1 shows the high-level metadata for

four typical research articles. Meetings with individual authors led to improvements in the granularity of the representation.

Topic	Method	Data	Evaluation	Research article title
Named entity recog.	HMM	Journal articles	Shared task	BioC Task1A: Finding NEs with a stochastic tagger
Summarization	Rule-based	Journal articles	Gold standard	Finding GeneRIFs via Gene Ontology annotations
Corpus	Distributional	Journal articles	Hypothesis testing	Text in traditional and Open Access scientific journals
Text classification	Rule-based	Journal articles	Gold standard	Classifying the contents of parentheses for text mining

Table 1: High-level annotations for four typical biomedical natural language processing papers.

The initial overlap after manual filtering of terminology extraction errors was 48%. Most missing concepts were very domain-specific, such as *electronic health record* and *biomedical text*.

Conclusion

Thanks to the combination of methodologies, this work has resulted in an ontology for the representation of natural language processing research articles that is both empirically supported by quantitative data, and vetted by members of the natural language processing community. In future work, we will continue to refine it; since our results showed low coverage of biomedical domain-specific NLP concepts, we will focus on that. Later we will develop tools for using it to index the biomedical natural language processing literature, and use the output of that indexing to explore the work's implications to enable reproducibility for NLP.

Acknowledgments

This research was supported [in part] by the Intramural Research Program of the NIH, NLM.

References

1. Hans Radder. *In and about the world: Philosophical studies of science and technology*. SUNY Press, 1996.
2. Antske Fokkens, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proc. Assoc. Comp. Ling.*, pages 1691–1701, 2013.
3. Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. Reproducibility of studies on text mining for citation screening in systematic reviews: evaluation and checklist. *J. Biomedical Informatics*, 73:1–13, 2017.
4. K Bretonnel Cohen, Jingbo Xia, et al. Three dimensions of reproducibility in natural language processing. In *Proc. Lang. Res. and Eval.*, page 156. NIH Public Access, 2018.
5. Mark D Wilkinson, Michel Dumontier, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, 2016.
6. Gretchen P Purcell and Edward H Shortliffe. Contextual models of clinical publications for enhancing retrieval from full-text databases. In *Proc. Ann. Symp. on Computer Application in Medical Care*, page 851, 1995.
7. David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7):e1000097, 2009.
8. Fan Wang, Richard L Schilsky, et al. Development and validation of a natural language processing tool to generate the CONSORT reporting checklist for randomized clinical trials. *JAMA Network Open*, 3(10), 2020.
9. Alvis Brazma, Pascal Hingamp, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.
10. Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.
11. T Hernandez-Boussard, S Bozkurt, et al. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *JAMIA*, 27(12):2011–2015, 2020.
12. Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36, 2014.